**Reasoning Strategies for Diagnostic Decision Making**

**Abstract**

Medical diagnoses are complex decisions that require seeking and integration of different sources of information on the part of doctors. Past work has found that confidence and accuracy in diagnoses can vary significantly even with more information, leading to diagnostic overconfidence that is especially impacted by increased medical experience. There has been scant work looking at information seeking patterns for medical diagnoses however, and how they might contribute to both confidence and accuracy. We recruited UK medical students (N = 85) in an online vignette-based diagnostic study, and Oxford medical students (N = 16) for a similar in-person think-aloud version of where students performed 6 diagnostic scenarios based on past patient cases. During each scenario, students sought information from the patient’s medical history, physical examination and testing. The online study found that diagnostic accuracy is associated with seeking more relevant information and with less variability in information seeking on a case-by-case basis. However, confidence, not accuracy, was associated with seeking more information. The think-aloud study, revealed that different reasoning strategies affect how diagnostic differentials are evaluated, in turn affecting information seeking during diagnostic decisions. A reanalysis of the online dataset using the reasoning strategies assigned in the think-aloud study revealed an interaction between strategy and the number of initial diagnoses with regards to accuracy, indicating clinicians should base their choice of reasoning strategy on their initial diagnostic uncertainty to maximise accuracy. Future work should focus on interventions for prompting suitable reasoning strategies and information seeking to improve diagnostic accuracy.

**Introduction**

**Diagnosis**

*“Problems in diagnosis have…been heavily dominated by physicians with little input from the cognitive sciences. What is missing…is foundational work aimed at understanding how clinicians in actual situations take a complex, tangled stream of phenomena…to create an understanding of them as a problem.” (Wears, 2014)*

Imagine a group of doctors within a hospital’s intensive/critical care unit. They are engaged in a collective discussion about a particular patient. The patient has presented with a series of symptoms, including dizziness, breathing difficulties and eventual chest pain. She has been placed under continuous monitoring of her ‘vital signs’, including heart rate, body temperature, blood pressure, blood oxygen saturation and respiration rate. There has been a slow decrease in her blood pressure and blood oxygen saturation. The doctors are deciding what is the most likely cause of this patient’s symptoms and how this may inform her future care/treatment. It is possible that the patient is suffering from pulmonary oedema, whereby fluid is collected in the air sacs of the lungs, causing severe and sometimes fatal congestion. The symptoms could also be suggestive of a tension pneumothorax, when a lung collapses. Alternatively, there could be a cardiac cause of the patient’s condition. The doctors must integrate the information they have so far, align their individual mental models of the patient and decide the following:

1. Do they have enough information to diagnose the patient’s condition?
2. If not, what extra information do they need? Are there further tests that need to be performed?
3. What actions should they start taking to treat the patient given the most likely diagnosis?

One of the difficulties within this scenario is that symptoms may be indicative of multiple underlying conditions. This example is illustrative of why many medical decisions are ‘ill-structured’ problems: they present several possible courses of action, and produce disagreements over both the current hypothesis for the patient’s condition and desired end goal for that patient’s care (Jonassen, 1997). Medical staff involved in a patient case can independently formulate very different understandings of a patient’s condition and how it would be best to proceed. They have to then align their thoughts in order to align their actions as a cohesive team.

Diagnosis is a core aspect of a doctor’s job and is important for a number of reasons. Firstly, accurate diagnosis is crucial to a patient’s treatment. Secondly, from a psychological standpoint, it allows for an extension of previous research on information gathering and confidence to an ecologically valid, real-world setting. Finally, past work looking at diagnosis has not yet provided clarity on the causes of diagnostic errors.

A report from the US Institute of Medicine (McGlynn, McDonald & Cassel, 2015) concluded that most patients will experience a diagnostic error within their lifetime. When looking at records of new diagnoses for spinal epidural abscess in the US Department of Veteran Affairs, Bhise et al. (2017) found that up to 55.5% of patients experienced diagnostic error. The Quality in Australian Health Care Study found that 20% of adverse events were due to delayed diagnosis (Wilson et al., 1999). Around 32% of clinical errors have been found to be caused by clinician assessment, particularly the clinician’s failure to weigh up competing diagnoses (Schiff et al., 2009). Even using the most conservative of these estimates, the scale of the diagnostic error is substantial when extrapolated to the population of patients. Diagnostic errors have also been found to lead to longer hospital stays and even increased patient mortality (Hautz et al., 2019).

Diagnostic error is by no means the sole cause of medical incidents. There are a number of factors tied to the wider work environment, culture and technology that can contribute to incidents and errors. A lot of these factors are challenging to isolate and emulate in an experimental setting. By understanding the individual psychological factors of the diagnostic process however, we better understand how sociotechnical and environmental factors interact with and amplify individual contributors to diagnostic error. Gaining a greater understanding of the causes of diagnostic error can have important implications for future interventions within healthcare settings.

**Cognitive Biases and Overconfidence in Diagnoses**

Diagnostic error can stem from cognitive biases during the diagnostic decision making process, such as primacy (Frotvedt et al., 2020) or recency (Chapman, Bergus & Elstein, 1996) biases. While it seems intuitive that classical decision making biases affect those in healthcare too (Restrepo et al., 2020), the empirical evidence of impact for medical decision making is scant, (van den Berge & Mamede, 2013). One example from dermatology looked found examples of satisficing bias (premature closure) and anchoring were found, but few examples of others such as availability and representative biases (Crowley et al., 2012). One type of bias that has manifested in more experimental findings is overconfidence (Berner & Graber, 2008, Meyer et al.., 2013).

At this point, we shall revisit the scenario presented at the start of this section. In summary, a patient is presenting with a set of symptoms, requiring doctors to assign a diagnosis to guide future treatment. One of the doctors confidently presents their opinion that the patient has suffered a pneumothorax.. The certainty with which the diagnosis is suggested makes it more difficult for others to disagree with, especially if the doctor is a consultant/attending such that there is a disparity in seniority.

Confidence can be viewed as one’s “subjective probability of a decision being correct” (Fleming & Daw, 2017). Confident individuals tend to be more influential on others in a group (Zarnoth & Sniezek, 1997) and can even causally increase the confidence of other observers (Cheng et al., 2021). This behaviour has been observed in mock jury trials, during which participants hear eyewitness testimonies presented with high confidence and then perceive those testimonies as more credible than testimonies provided with low confidence (Cutler, Penrod & Dexter, 1989, Roediger, Wixted & DeSoto, 2012). Confidence is a commonly used predictor of another person’s accuracy, especially when feedback is not readily available of an individual’s true accuracy. Confidence also varies across individuals with what may be considered a ‘subjective fingerprint’ (Ais et al., 2016), and individuals may be systematically underconfident or overconfident. Confidence has been explained computationally as the difference in the strength of evidence for a decision alternative compared to other alternatives (Vickers & Packer, 1982). After a decision is made, we continue to process evidence, i.e. we continue to think about a decision after it has been made and having ‘second thoughts’ or changes of mind are more likely with a lower level of confidence (Resulaj et al., 2009).

Individuals are ‘well-calibrated’ with regards to confidence if their internal likelihood of being correct is predictive of their true accuracy. However, confidence can become decoupled from true accuracy. This decoupling is known as ‘miscalibration’. One would show miscalibration of confidence if they tended to be more confident than they are correct (overconfidence) or more uncertain than they are correct (underconfidence).

In a task that involved diagnosing ultrasound scans, it was found that overconfidence was inversely associated with the amount of clinical experience that the clinicians/participants had (Schoenherr, Waechter & Millington, 2018). However, it has also been found that underconfidence can be more prevalent than overconfidence, especially when comparing medical students to residents (Friedman et al., 2005). Similarly, Yang and Thompson (2010) found that experienced nurses exhibited similar performance to nursing students, but were more confident in their judgements, showing differences in confidence calibration across experience levels. More broadly, highly confident members within a group could unknowingly reduce the chance of less confident members speaking up about potential errors, which is a common problem within healthcare (Hémon et al., 2020). Overconfidence has also been linked to a lower likelihood of sufficient patient management and clinical effort as per a field study in Senegal (Kovacs, Lagarde & Cairns, 2019).

We would argue that building on the current research landscape of diagnostic confidence is important. If there is an assumption that others will calibrate their confidence to their true accuracy, this would mean that heeding high confidence advice or judgements would be an optimal strategy for maximising accuracy. However, this can be a serious issue when high confidence errors lead others astray. This is important, as in addition to seniority and speciality experience, a clinician’s confidence is one of the only markers available for other clinicians and for patients when making key medical decisions. One underexplored avenue in current research is the role that information seeking during the diagnostic process affects confidence.

**Information Seeking**

Clinicians generate hypotheses and then gather information to reduce the space of hypotheses. They should ideally eliminate hypotheses from consideration only when it makes sense given the incoming evidence. By the same token, they should also not continue attaching themselves to a hypothesis when there is overwhelming evidence to the contrary. One conclusion of Wason (1960) was that individuals struggle to remove a hypothesis from consideration even if they receive evidence against it. Understanding how individuals generally reason about a possible space of hypotheses is interesting for understanding how the reasoning process works differentially for novices and experts, especially in a specialised domain such as medicine. One question that is worth investigating is how the ‘process of elimination’ affects confidence.

The link between confidence and information seeking has been previously investigated in cognitive psychology research. Information can be gathered that is either in support of or against an individual’s beliefs or decisions, with information being used to accumulate strength of evidence in favour of different decision alternatives (Vickers & Packer, 1982). Desender, Boldt & Yeung (2018) found that higher variability was associated with lower confidence and higher information seeking. However, the mere quantity of information, even if that information favours the non-preferred option, may increase confidence in of itself (Ko, Feuerriegel, et al., 2022).

There is also evidence to assume that information seeking is important within medical diagnoses too. Notably, Gruppen, Wolf & Billi (1991) found that clinicians were less confident when they had to seek relevant information for themselves compared to all information was already provided, indicating that information seeking as a task is contributory to formulating diagnostic confidence. While this shows the relationship in one direction, past work has also viewed confidence as contributory to further information seeking. Pathologists with more calibrated confidence were found to request more information, such as second opinions or ancillary tests, when unconfident in their judgements (Clayton et al., 2022). In a sample of 118 physicians presented with patient vignettes, it was found that higher confidence was associated with a decreased amount of diagnostic tests being ordered, even if confidence and accuracy were larger decoupled/miscalibrated (Meyer et al., 2013). It has also been observed previously that physicians may ‘distort’ neutral or inconclusive evidence to be interpreted as supporting prior beliefs (Kostopolou et al., 2012). Similarly, it has been found that a patient’s case history that suggests a particular diagnosis prompts selective interpretation of clinical features that favour the initial diagnosis (Leblanc, Brooks & Norman, 2002). Together, these findings have implications for how clinicians may seek and integrate evidence when making decisions and how patterns of receiving information could affect decision confidence and in turn confidence calibration.

Diagnostic decisions have been thought of as ‘ideal’ when using the hypothetico-deductive process (Kuipers & Kassirer, 1984), whereby hypotheses are formulated based on specific features of a patient and are then linked to established criteria for a diagnosis, with further information gathering to test these hypotheses (Higgs et al., 2008) or eliminate others. This account was challenged by Coderre et al. (2003), who found that diagnosis can based more on pattern recognition, especially for more experienced clinicians. Either way, the bridge between confidence and information seeking is the reasoning strategy utilised by clinicians. Diagnostic reasoning is currently taught using cognitive frameworks such as the surgical sieve and the ABCDE pneumonic. However, current education does not account for differences in reasoning strategies, whether strategies may meaningful vary by case and by clinician and how these strategies have a downstream influence on the diagnostic process in terms of seeking information, generating differentials and formulating confidence.

**Current Work**

There is a need for the teaching and assessment of non-technical skills and human factors in healthcare (Higham et al., 2019), which is currently not addressed in a widespread standardised manner in speciality curricula (Grieg, Higham & Vaux, 2015). Curricula within medicine also place little emphasis on how uncertainty is communicated and approached in medical decision making (Hall, 2002). In addition, there is little work that informs how information seeking is taught within medical reasoning other the use of cognitive frameworks (such as the ‘surgical sieve’) and pneumonics (such as Airway, Breathing, Circulation, Disability, Exposure). Clinical experience may also be connected to risk aversion and further information seeking behaviour (Lawton et al., 2019), which offers an important avenue for future medical education. Hence, this research informs medical education of non-technical skills such as diagnostic reasoning, especially around evaluating diagnostic differentials and seeking information during the diagnosis process.

Here, we present two studies that use a similar vignette-based experimental paradigm. The first study is with a larger sample using online data collection to characterise information seeking and confidence within diagnostic differential generation. We follow this up with a smaller sample using a think-aloud protocol to gain insight on diagnostic reasoning strategies using the work of Coderre et al. (2003). Across these two studies, we aim to investigate whether patterns of information seeking and reasoning strategy predict diagnostic accuracy and confidence. Dependent on whether this is the case, we characterise which aspects of information seeking predict each of these, be it seeking more information, better information or the suitability of information to reasoning strategy. During both of these tasks, we allow medical students to generate multiple diagnostic differentials to better reflect their uncertainty in not settling on a single diagnosis. Our paradigm also allows for flexible information gathering to emulate naturalistic seeking patterns during diagnostic decisions, up until the point where a clinician decides that they are ready to treat a patient. We finally present an re-analysis of the online dataset in light of the think-aloud study to investigate whether reasoning strategy is predictive of diagnostic accuracy.

**Study 1 – Diagnostic Study**

We sought in this study to better understand how information seeking, confidence and differential generation interact within the diagnosis process. In particular, we investigated whether information seeking patterns were associated with diagnostic accuracy and confidence. We conducted a vignette-based diagnosis study with medical students to inform future work on how diagnostic reasoning is taught to students, especially when it comes to weighing up competing differentials. Data is openly available on OSF: <https://osf.io/kb54u/>.

**Methods**

**Participants**

We recruited medical students within the UK who were in the year of study before their foundation programme. 85 medical students completed the study, including 32 males, 52 females and 1 participant who reported as non-binary. Their ages ranged between 22-34 years (M = 24.2). Participants were recruited between July 11th 2022 and April 6th 2023. The UK Medical Schools Council distributed the study to UK medical students using a mailing list. Participants were emailed with a study information sheet and a link to access the experiment, where they first provided consent via an anonymous online form. After doing so, the participant provided demographic information (age, gender and years of medical experience). The study was conducted online, with participants able to run the experiment in a browser on a desktop computer or laptop. Participants were not able to run the experiment on their phone or tablet. The experiment was coded using the JSPsych Javascript plugin. The code is publicly available on Github: <https://github.com/raj925/DiagnosisParadigm>. This study was reviewed and granted ethical approval by the Oxford Medical Sciences Interdivisional Research Ethics Committee under reference R81158/RE001. Informed consent was obtained anonymously using an online electronic information sheet and consent form.

**Materials**

This study involved patient vignettes that have been adapted from actual past cases. We adapted scenarios from a bank of patient cases from Friedman (2004). Our study involved 6 patient cases, each designed to indicate a specific underlying condition that the real patient had. These conditions were: Aortic Dissection (AD), Guillain-Barre Syndrome (GBS), Miliary TB (MTB), Temporal Arteritis (TA), Thrombotic Thrombocytopenic Purpura (TTP) and Ulcerative Colitis (UC). The order in which the cases were presented was randomised for each participant. We also included a practice case (Colon Cancer) to familiarise the participants with the experimental procedure and the interface.

A panel of 3 subject matter experts (practising medical staff and researchers within the NHS) who work in collaboration with the OxSTaR Centre (Oxford Simulation, teaching and Research Centre, based at Oxford’s John Radcliffe Hospital) were recruited to design the vignettes used in this study. These medical professionals were at differing experience levels, with their medical roles at the time of this study as follows: Speciality trainee (ST6) in Anaesthetics, Foundation (F1) Doctor and Gastroenterology Consultant. The panel assisted with translating terms (e.g., medication names, tests etc.) from US to UK doctors’ vernacular and updated patient details to be more current. The panel also crucially gave input on which medical conditions to choose from the provided bank of vignettes that medical students would be expected to know, whilst having some variance in both difficulty and the affected pathophysiological system.

**Procedure**

The goal of the task was to determine a diagnosis, or set of diagnoses, for each presented patient (Figure 1). Information on the patient is split into a series of discrete stages so that the researchers are able to control what information the clinicians have access to at any given point in the experiment. We can call each point of new information an “information stage”. Participants are able to seek information freely until they are ready to move on, similar to the paradigm adopted by Kämmer et al. (2019).

A diagram of a patient's flow

Description automatically generated

*Figure 1: Paradigm of Study 1, showing the procedure for a single patient case. For Study 2, the procedure was very similar but excluded the ‘Indicate Differentials’ stage.*

The procedure of a single case is as follows. The participant is asked to imagine that they are working in a busy district hospital and they encounter patients in a similar way to how they would in their real medical practice. At the start of each case, the participant is shown a description of a patient, which includes the patient’s gender, age and their presenting complaint. An example of this is: “patient is a 68 year old male presenting with fever and arthralgia”. Each case is split into three information stages: Patient History, Physical Examination and Testing (in this order). The set of information requests for each stage is the same for all cases. The Patient History stage includes information on “Allergies”, “History of the Presenting Complaint”, “Past Medical History” and “Family History”. The Physical Examination stage includes ‘actions’ that a doctor may take when examining a patient, such as “auscultate the lungs”, “abdomen examination”, “take pulse” and “measure temperature”. Finally, the Testing stage involves information on any bedside tests or tests they may request from another department. This includes “Chest X-Ray”, “Venous Blood Gas”, “Urine Dipstick” and “Clotting Test”. In total, 29 possible tests that can be requested across the three information stages.

A screenshot of a computer

Description automatically generated

*Figure 2: Screenshot of the interface. Shown here is the screen in which the participants seek information during the Testing stage.*

When a participant clicks on any of these tests, the information for that test is shown on screen after a 3 second delay. It was emphasised during the task instructions that participants should only request information that they believe will help them with diagnosing the patient for that specific case. Participants are free to request the same piece of information multiple times, including information from a previous stage. At any point, they can choose to stop gathering information for that stage. They are then taken to a new screen where they report a list of all differential diagnoses that they are considering for that patient at that stage. For each differential, participants report a “level of concern” for that differential, which is how concerned they would be for that patient if this differential really was the patient’s underlying condition. This is reported on a 4 point scale, with labels of “Low”, “Medium”, “High” and “Emergency”. Participants also reported a likelihood rating for each differential, ranging from 1 (very unlikely) to 10 (certain). In subsequent stages, the list from the previous stages is available for participants to update concern/likelihood ratings, or to add/remove differentials from the list. Even at the last information stage, participants can report multiple differentials.

A screenshot of a computer screen

Description automatically generated

*Figure 3: Screenshot of the interface. This is the screen in which participants report their current list of differentials, including the name of each condition as well as the severity and likelihood ratings for each condition. Participants remove conditions by clicking the red cross on the right hand side of each differential. Participants add a new differential by clicking the plus icon below the list.*

After recording their differentials, participants are then asked to report their confidence that they are “ready to start treating the patient” on a 100 point scale, ranging from fully unconfident to fully confident. Participants also indicate using a checkbox whether they are ready to start treating the patient, at which point a text box appears for them to report what further tests they would perform, any escalations they would make to other medical staff and treatments they would start administering for the patient. Once all three stages are complete, participants report how difficult they found it to determine a diagnosis for that case, on a scale from 1 (trivial) to 10 (impossible). At the end of all six patient cases, participants are told the ‘true’ conditions for all the patients.

**Data Analysis**

Responses were coded for correctness manually with help from a medical consultant, who looked at all the information available for each case and determined which diagnoses could be valid answers. All lists of differentials were ‘marked’ for correctness manually using the criteria found in Table S1 of the Supplemental Materials.

We test for correlations between our dependent variables using Pearson’s product moment correlation tests, and interpret an alpha value of less than 0.05 as indicative of a statistically significant effect. Our sample of 85 participants is calculated have 80.4% power to detect a medium effect size of r = 0.3 (using an approximate arctangh transformation correlation power calculation). Our key dependent variables are as follows:

Case-Wise Measures

Accuracy

For a case to be considered ‘correct’, the participant should have reported the correct condition for that case within their list of differentials regardless of the number of differentials provided. Given that differentials are provided via free text, cases are manually coded as correct or incorrect using the aforementioned criteria. Our main accuracy measure is computed by the taking the likelihood value assigned to the correct differential if it is included in the list of differentials. This means that likelihoods range from 1-10 when a correct differential is included and has a value of 0 when a correct differentials is not included. The value is then rescaled to range from 0 and 1, where 1 corresponds to a correct differential assigned maximum likelihood. If multiple differentials that are considered correct were provided, then the likelihood value of closest differential to the true condition was used.

Confidence

Participants reported confidence at each information stage. Initial Confidence refers to the reported confidence after the first stage of information seeking (Patient History), whilst Final Confidence refers to the reported confidence after the third and last stage of information seeking (Testing). As for accuracy, confidence is rescaled to fall between 0 and 1 to allow for direct comparison between the two variables. We can then use these two variables to calculate Confidence Change, by subtracting the participants' Initial Confidence from their Final Confidence. Hence, a positive value for Confidence Change means that the participant has gained confidence over the course of the patient case.

Number of Differentials

We record the number of items in the list of differentials at each stage. Initial Differentials refer to the number of differentials after the first stage of information seeking (Patient History), whilst Final Differentials refer to the number of differentials after the third and last stage of information seeking (Testing).

Perceived Difficulty

The subjective rating by participants at the end of each case for how difficult they found it to determine a diagnosis for that patient case. This is reported subjectively by each participant on a scale from 1 (trivial) to 10 (impossible).

Derived Information Seeking Measures Across Cases

Amount of Information Seeking

We take the number of unique tests requested at a given information stage (i.e. not including any tests from a previous stage or including tests that had been requested before during that stage) and divide this by the number of possible tests available.

Information Seeking Value

We compute the average value of sought information across cases. To do this, we take each of the 29 pieces of information in turn by case and split trials into two groups: trials of that case where that information was sought and trials of that case where that information was not sought. For each group, we compute the proportion of trials where the students included a correct differential, and then take the difference between these two values. A positive value would indicate that students were more likely to identify the correct condition with that information rather than without that information. This difference can be considered that information’s ‘value’. For each of the participants’ cases, we compute this difference for each piece of information that the participant sought (for information they did not seek, the informational ‘value’ would be 0) and then calculate the sum of all information values for each case. We then take the mean information value across all cases for each participant. This gives an overall measure of, on average, how useful the information was that participants sought on each case.

**Results**

Overall Performance

Across cases, the proportion of trials where participants include a correct differential within their set of differentials increased with each stage of information gathering. (F(2, 128) = 59.52, η2G = .08, p < .001). Participants included a correct differential on fewer trials during the Patient History stage (M = 0.54, SD = 0.23) than during the Physical Examination (M = 0.66, SD = 0.22) and Testing stages (M = 0.69, SD = 0.21). Table 2 shows overall performance by case, indicating that there was variability in performance due to cases varying in difficulty.

Calibration of Confidence to Accuracy

Confidence also increased as participants received more information (F(1, 123) = 75.45, η2G = .15, p < .001). Participants reported lower confidence during the Patient History stage (M = 0.30, SD = 0.15) than during the Physical Examination (M = 0.41, SD = 0.17) and Testing stages (M = 0.47, SD = 0.19). We note here that confidence was on average below 50% even at the end of each case, which may indicate that participants were not highly confident to start treatment.

When comparing Accuracy (taking into the likelihood assigned to correct differentials) to Confidence, we find, across stages, participants’ Confidence was fairly well aligned to their Accuracy (see Figure 4). To determine whether confident participants tended to be more accurate, we compared a paired t-test between Average Confidence and Average Accuracy (across cases) at each stage. We did not evidence of a difference between the two at the Patient History (t(84) = 0.32, MDiff = 0.01, p = .75) and Physical Examination stages (t(84) = 0.75, MDiff = 0.01, p = .45), but we do see evidence of a difference between the two at the Testing stage (t(84) = 2.40, MDiff = 0.06, p = .02). This indicated well-calibrated confidence after Patient History and Physical Examination, but a slight overconfidence across participants after Testing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Case** | **Proportion of Participants who Included a Correct Differential** | **Accuracy** | **Perceived Difficulty** | **Mean Final Confidence** |
| UC | 0.92 | 0.73 | 5.30 | 0.61 |
| GBS | 0.69 | 0.54 | 6.88 | 0.37 |
| TA | 0.66 | 0.67 | 6.16 | 0.48 |
| TTP | 0.55 | 0.55 | 6.80 | 0.41 |
| AD | 0.53 | 0.47 | 5.92 | 0.49 |
| MTB | 0.42 | 0.57 | 6.69 | 0.45 |

*Table 2: Showing statistics across participants for each case (leftmost column). Differential Accuracy refers to proportion of cases in which a correct differential is included in the list of differentials, whilst Accuracy refers to the average likelihood (on a 1-10 scale) assigned to a correct differential if included. Both of these measures, as well as Final Confidence, are calculated at the final information stage of each case (i.e. the Testing stage).*

A graph showing a line of a performance

Description automatically generated with medium confidence

*Figure 4: Graph showing Accuracy (black) and Confidence (green) at each of the three information stages.*

Differentials

We first look at the number of differentials that participants report at each stage. Participants overall increased the number of the differentials they reported as they received more information (F(1, 107) = 94.02, η2G = .08, p < .001). Participants reported fewer differentials during the Patient History stage (M = 3.20, SD = 1.11) than during the Physical Examination (M = 3.88, SD = 1.33) and Testing stages (M = 4.12, SD = 1.43). The majority (74/85) did not decrease the number of differentials between Patient History and Testing on any case.

To look at whether the number of initial differentials generated the amount of information sought, we conducted a Pearson’s Correlation test on individual differences. We find an association (see Figure 7) between the average number of differentials generated from the Patient History and the average amount of information sought during cases (r(83) = 0.30, 95% CI = [.10, .49], p = .005). As previously discussed, participants rarely seem to remove differentials from consideration. Therefore, one can surmise here that higher information seeking is associated with the consideration of more diagnostic differentials.

A graph of a number of differentials

Description automatically generated

*Figure 5: Scatter plot showing the relationship between the number of initial differentials (x-axis) and the proportion of available information sought (y-axis). Each point represents a single student with both variables average across the six cases that each student performs. The x-axis refers to the average number of differentials that participants report in their list at the Patient History stage. The y-axis refers to the average proportion of available information sought, with each case containing 29 pieces of information across the Patient History, Physical Examination and Testing stages. The line of best fit is plotted using the geom\_smooth function in R with a linear model. The shaded region shows the 95% confidence interval of the correlation.*

Information Seeking

The Proportion of Information Seeking decreased with each information stage (F(2, 151) = 122.0, η2G = .30, p < .001). Participants sought more of the available information during the Patient History stage (M = 0.85, SD = 0.20) than during both during the Physical Examination (M = 0.59, SD = 0.24) and Testing stages (M = 0.50, SD = 0.22).

When conducting a Pearson’s Product Correlation test, we do not find that participants who sought more information across cases were also more accurate in their diagnoses (r(83) = 0.17, 95% CI = [-.04, .37], p = .11). However, participants who sought more information did tend to increase their confidence more (r(83) = 0.24, 95% CI = [.02, .43], p = .03, see Figure 5). This is distinct from their final confidence, for which we did not find evidence of an association with the amount of information sought (r(83) = 0.11, 95% CI = [-.11, .31], p = .33). While seeking more information may imbue students with a greater level of confidence, it does not necessarily translate into more accurate diagnoses. This links to the results presented in Figure 4, in which confidence and accuracy were related to one another but imperfectly (especially during the Testing stage).

A group of graphs with numbers and symbols

Description automatically generated with medium confidence

*Figure 6: Scatter plots showing our information seeking variables (amount and value) against our key dependent variables of change in confidence (difference between final confidence and initial confidence) and accuracy (which is divided by 10 in this case so that values fall between 0 and 1). Information Amount refers to the proportion of available information sought across cases. Information Value refers to the sum of all mean information values across all 6 cases for a given participant. All data points are for a single participant where variables are averaged across all 6 cases they completed.*

The amount of information sought does not seem to be predictive of accuracy. However, it may be the case that patterns of information sought are instead predictive of differences in accuracy on this task. In order to test this, we investigate whether information seeking is predictive of participants who are higher or lower in their diagnostic accuracy using binary classification and receiver operating characteristic (ROC) analysis. We trained a binary classification algorithm using a generalised logistic regression model to identify if participants of high or low accuracy based on the information they sought. We split all cases by whether they performed by a high and low Accuracy participant using a median split of participants by their average Accuracy across the six cases. We train the classifier using a Generalised Linear Model (GLM) by treating the 29 binary variables for each information as predictors (with a 1 signifying that the information was sought for that case and 0 when the information was not sought) to predict the binary outcome of whether the participant is a low or high accuracy participant. We used Leave One Out Cross Validation, such that each case is predicted by training the algorithm on all other cases. By plotting an ROC curve of our classifier, we find an area under the curve (AUC) value of 0.72 (plotted in Figure 7). When conducting a DeLong test, to test the null hypothesis that the AUC is equal is 0.5 (i.e. that the classifier is completely unable to predict high and low accuracy participants), we find p < .001, indicating that the AUC differs significantly from 0.5 and that the classifier is able to reliably predict high and low accuracy participants.

This indicates overall that differences in information seeking are indeed predictive of a difference in participant ability at above chance at a broad level. Essentially, information seeking patterns separate high and low accuracy participants, but this analysis does not tell us what aspects of information seeking in particular are predictive of accuracy. We next seek to identify and better characterise these specific differences in information seeking that contribute to this relationship with diagnostic ability by correlating behavioural information seeking variables with accuracy.

A graph of a curve

Description automatically generated

*Figure 7: Receiver-Operator Characteristic (ROC) curve using a Generalised Linear Model to classify individual cases as being performed by either high or low accuracy participants. The models are trained on the raw binary predictor variables for each of the 29 available pieces of information, with 0 indicating that the information was not sought for the case and 1 indicating that the information was sought. Participants were sorted as high and low accuracy based on a median split on their average Accuracy value across the six cases.*

In order to examine more specifically what differences in information seeking are driving differences in both accuracy and confidence, we look at the relationship between accuracy and informational value. We assess the degree to which each participant’s accuracy is predicted by the quality of the information they sought and find evidence for a positive relationship between accuracy and information value (r(83) = 0.25, 95% CI = [.04, .44], p = .02), as well as between confidence and information value (r(83) = 0.28, 95% CI = [.07, .47], p = .01).

**Discussion**

This study of 85 medical students found that accuracy and confidence are well calibrated. Previous work (e.g. Meyer et al., 2011) have noted a gap between subjective confidence and objective accuracy. In particular, there has been demonstrated to be a general tendency for less experienced medical trainees to be underconfident and for more experienced medical professionals to be overconfident (Yang and Thompson, 2010). Part of this discrepancy between our findings and past findings could stem from the diagnostic uncertainty expressed by students in this study, which they do in two ways. Firstly, students broaden, rather than narrow, their considered diagnostic differentials with more information and still report a broad range of differentials after receiving all available information for a given case. There is a general adage in healthcare that medical students come across which says that “history is 80% of the diagnosis”. It is therefore worth considering whether there is a specific facet of diagnostic decisions whereby students are taught not to disregard diagnostic possibilities easily. Secondly, students reported fairly low confidence overall to treat patients, with an average confidence of below 50% even after receiving all available information. This may indicate that part of ensuring appropriate confidence, or expressions of uncertainty could be related to properly evaluating all possible diagnostic differentials rather than forcing decisions to focus on a single diagnosis, which has been cited previously as a problematic tendency (Redelmeier & Shafir, 2023).

We find the amount of information sought informed confidence, whilst accuracy was associated with seeking more useful information on each case. This hints at the richness of this dataset in picking on information seeking and differential generation behaviour. We note however that whilst predictors of diagnostic by information seeking behaviour were found, they do not tell us how overarching differences in such behaviour arise. One possibility is that these differences stem from reasoning strategies that we cannot infer from this current dataset. In order to ascertain these strategies, we conduct a follow-up study using a similar diagnostic paradigm conducted in-person where students think out loud as they make diagnoses. We use criteria taken from Coderre et al. (2003) to code case by the reasoning strategy employed. We hypothesise that different reasoning strategies for generating differentials are useful for some cases more than others and that information seeking varies as a function of strategy. This coding of reasoning strategies is then subsequently used to classify the same reasoning strategies in the online dataset from study 1 (where we do not have access to the participants’ thought process) by using the information they sought.

**Study 2 – Think-Aloud Study on Diagnostic Reasoning Strategies**

We aimed to replicate the finding of considered differentials increasing with more information when the method by which these differentials were reported. Are students seeking information to confirm their existing set of differentials, to rule out differentials or to expand their set of considered possibilities? And are these different approaches interleaving or are they more dependent on individual diagnostic decision making styles? In order to provide more context to the results from study 1, we conducted a follow-up study that utilised a very similar experimental procedure, but instead prompted students to think out loud as they were performing the task. and the transcripts were coded to conduct both quantitative and qualitative analysis.

Think-aloud methodologies are useful for directly accessing ongoing thought processes during decisions (van Someren, Barnard & Sandberg, 1994). The use of thinking aloud (or verbal protocols) in research is useful for being able to access the information attended to participants in short term memory (Payne, 1994) and can be treated as the ongoing behavioural state of a participant’s knowledge (Newell & Simon, 1972). Think-aloud protocols have historically been used to study problem solving, particularly for comparing how novices and experts solve problems such as finding the best move in chess (de Groot, 1946, Bilalić, McLeod & Gobet, 2008). Diagnosis is a decisional process that develops over time and allowing participants to think aloud reflects this by providing a time-ordered sequence of how thought processes develop (Payne, 1994). This is especially well-suited to our task where the information available to participants is controlled with time, allowing us to investigate how diagnostic thinking develops with more information. A think-aloud methodology has previously been used to study the differences between novice and expert clinicians during diagnostic reasoning (Coderre et al., 2003). This study found a general trend that experts tend to use a ‘pattern recognition’ approach to diagnosis more than novices, who tended to use a ‘hypothetico-deductive’ process (which is aforementioned to be the ‘textbook’ description of the diagnostic process), but this was highly dependent on the case presented. We build on the work of Coderre et al. (2003) here to further investigate how reasoning strategies contribute to accuracy and why certain cases result in differing strategies.

**Methods**

**Participants**

16 participants were recruited for this study. Participants were 5th or 6th year medical students at Oxford University (including 2nd year Oxford University Graduate Entry Medical students) recruited via posters in the John Radcliffe Hospital in Oxford and via a mailing list for students managed by the Medical Sciences Division at the University of Oxford. The study was conducted onsite at John Radcliffe hospital. Participants were recruited between July 5th 2023 and December 1st 2023. Data was reviewed on an ongoing basis to cease recruitment when emerging themes were exhausted. This study was reviewed and granted ethical approval as an amendment to our existing protocol to allow for audio recordings by the Oxford Medical Sciences Interdivisional Research Ethics Committee under reference R81158/RE004.

**Materials**

The same set of cases and a similar computer interface from Study 1 were used for this study, with the exception that participants no longer recorded their differentials in a specific screen at the end of each information gathering stage. Instead, participants’ differentials were recorded as a more naturalistic part of their diagnostic process as they reported aloud their thoughts as they worked through each diagnostic case. The study was conducted onsite using a laptop, with actions on screen recorded on video and the audio of participants’ thinking aloud recorded via a microphone. Informed consent was obtained anonymously using an online electronic information sheet and consent form. Information, including experimental data and audio recordings, collected during the study were stored under anonymised IDs with no linkages to participants. Data was kept on a password-protected computer and hard drive.

**Procedure**

The general procedure was very similar to that of Study 1, except that participants were given the following instructions at the start of the study:

*“Whilst you are doing the task, you will be asked to think aloud. This means that you verbalise what you are thinking about, especially how you interpret the information you receive and what conditions or diagnoses you are considering or are concerned about for each patient case. If you have nothing to say or nothing on your mind, there’s no need to say anything but do say whatever is on your mind once it pops up. If you are unsure about anything you see or do not know about what something means, you will not receive any help but verbalise when you are unsure about anything during the task. Please make sure that you speak clearly ‘to the room’.”*

The experimenter occasionally prompted participants with content-neutral probes: “can you tell me what you are thinking?” in cases of periods of long silence, and “can you tell me more?” when the participant said something vague that may warrant further detail. We emphasise that these are non-leading questions. The audio of the participants’ verbalisations was recorded and then transcribed. An initial transcript was generated using Microsoft Office’s transcription feature, but the transcript was checked and modified for accuracy by listening through the audio recordings again. The screen of the experimental interface was also recorded, such that the audio could be linked to specific actions within the task. The focus of this study is on verbal utterances rather than any non-verbal or inferential aspects of the participants’ qualitative data. At the end of the experiment, the researcher administered a semi-structured interview to better understand what the participants feel their diagnostic reasoning approach tends to be. These questions are provided in the Supplemental Materials.

**Data Analysis**

We conducted a theory-driven semantic thematic analysis (as per definitions detailed by Braun and Clarke, 2006) to code utterances under specific categories. This kind of thematic analysis is suitable given that our qualitative data is from a structured experiment, rather than a dataset with a looser structure (e.g. interview recordings). As a result, we apply deductive analysis using predetermined codes for think-aloud utterances and for a debrief interview where we administer a semi-structured interview with specific questions of interest.

Firstly, we code all utterances related to the main research areas of interest in this project, namely information seeking, confidence and differential/hypothesis generation. Respectively, we define the following codes:

* **Differential Evaluation:** any time that the participant (each of the following is considered a separate subcode):
  + Differential Added - Mentions a new condition that they are considering
  + Differential Removed - Rules out or eliminates a condition from consideration
  + Likelihood Increased - Mention of increased likelihood of a previously mentioned condition, or that information seems to correspond with a condition
  + Likelihood Decreased - Mention of decreased likelihood of a previously mentioned condition, or that information seems to contradict with a condition
* **Information Seeking Strategies:** any time the participant expresses why they may or may not request a particular piece of information in relation to ruling out or confirming a condition.

We also define a group of codes that indicate three different diagnostic reasoning strategies: hypothetico-deductive reasoning, scheme-inductive reasoning and pattern recognition (Coderre et al.., 2003). These were defined as follows:

* **Hypothetico-Deductive Reasoning** - prior to selecting the most likely diagnosis, the participant analysed any alternative differentials one by one through something akin to a process of elimination.
* **Scheme Inductive Reasoning** - participant structures their diagnosis by pathophysiological systems or categories of conditions (e.g., infective vs cardiovascular causes) to determine root causes of patient symptoms rather than focusing on specific conditions.
* **Pattern Recognition** - participant considers only a single diagnosis with only perfunctory attention to the alternatives, or makes reference to pattern matching when using a prototypical condition to match its symptoms against the current observed symptoms for the patient (e.g., “these symptoms sound like X” or “this fits with a picture of Y”).

We first code specific statements within each case that suggested one of these strategies, and then determined which strategy was most prevalent or influential for cases as a whole such that each case was categorised under one of these strategies. In addition to coding each case under one of these strategies, we also code participants on an overall level based on their subjective perception of how they make diagnostic decisions. This is based on responses provided during the debrief interview (as described in the Procedure section). Hence, reasoning strategy codes are at the case level and also at the participant level.

Coding of utterances and case-wise reasoning strategies were conducted with a second independent coder. For reasoning strategies, initial interrater reliability was low, with both coders agreeing on 58.3% of cases. Conflict resolution led to changes made to the coding criteria by prioritising strategies used early in a case, as some participants were noted to utilise multiple strategies within a single case, as well as allowing some cases to be coded as not having a clear strategy due to a lack of utterances. Conflicts were then resolved with these updated criteria. Both coders agreed on 78% of cases when coding for correctness, with conflicts resolved in consultation with a member of expert panel used to develop the vignettes (as mentioned in Study 1).

Although we do not record differentials in the same way as in Study 1 (in a list with corresponding likelihood and severity ratings), we do obtain the other variables from Study 1. Namely, we record confidence at each stage of information seeking and data around the information sought by participants. As we do not explicitly record differentials in the same manner as in Study 1, accuracy is operationalised differently. We code each case as ‘correct’ if a correct differential is mentioned at some point by the participant (using the same marking scheme in table S1).

**Results**

First, we look at overall quantitative characteristics of the think aloud statements. When looking at accuracy (the proportion of cases where a correct differential was mentioned by the participant), accuracy was 0.57 across all cases. This varied considerably by condition however, with accuracy across participants for each condition being as follows: AD = 0.63, GBS = 0.88, MTB = 0.19, TA = 0.44, TTP = 0.69, UC = 0.63. For utterances coded as Differential Evaluations, participants on average made 5.21 such utterances per case (SD = 2.80). The mean number of Differential Evaluations was relatively constant by condition except for the AD case: AD = 8.18, GBS = 4.63, MTB = 4.81, TA = 4.75, TTP = 4.25, UC = 4.63. Participants varied in how much they spoke during the study, uttering 1038-7730 words (M = 4194) across the scenarios. Part of this range is driven by participants repeating information they see during the task, but participants also varied in terms of how much they externalised their thought process.

A graph of different colored bars

Description automatically generated

*Figure 8: Accuracy for each case in the think-aloud study across participants. Accuracy is operationalised here as the proportion of participants who mention a differential considered correct during that case. Cases are ordered on the x-axis in decreasing order of accuracy.*

As previously mentioned, Differential Evaluations can be further categorised into one of four subcodes: Differential Added, Differential Removed, Likelihood Increased and Likelihood Decreased. As found in the previous study, there is a general reticence to disregard differentials completely. Participants expressed significantly more statements adding differentials (M = 3.14, SD = 0.89) than removing differentials (M = 0.27, SD = 0.28) (t(15) = 14.14, MDiff = 2.86, p < .001). Participants expressed more statements of decreasing likelihoods (M = 0.99, SD = 0.62) rather than increasing likelihoods (M = 0.93, SD = 0.46) but we did not find evidence of a significant difference (t(15) = 0.34, MDiff = 0.06, p = .73).

A comparison of different colored squares

Description automatically generated with medium confidence

*Figure 9: Bar graphs showing the average number of each type of differential evaluation across all cases by all participants. The figure on the left (9a) compares the number of statements describing new differentials being considered (green) against the number where a differential was eliminated or placed not under consideration (purple). The figure on the right (9b) compares the number of utterances where a differential is decreased in likelihood or is said to have contradicting information (light blue) against the number of utterances where a differential is increased in likelihood or is said to have supporting information (brown).*

Reasoning Strategies

Next we look at our coding of reasoning strategies at a case level. As mentioned, our criteria for each code was applied to each individual case based on the transcribed utterances. When looking at reasoning strategies by case, 43 cases were coded as Hypothetico-Deductive, 29 were coded as Pattern Recognition and 18 were coded as Scheme Inductive (the remainder of cases did not contain enough clear utterances to classify under one of these strategies). Accuracy was higher for cases coded as Hypothetico-Deductive (71%) compared to both Pattern Recognition cases (64%) and Scheme Inductive (39%). It is worth noting here that accuracy was solely based on participants mentioning differentials during their thinking aloud, which is naturally not facilitated by Scheme Inductive reasoning due to its focus on identifying pathophysiological systems acting as sources of patient symptoms rather than specific conditions. This can hence explain the lower ‘accuracy’ for Scheme Inductive cases. We also note that the types of reasoning strategy used varies by condition (see Figure 13 below), with the MTB and TTP cases in particular exhibiting higher usage of Pattern Recognition than others. This could be because this case was considered harder than others and hence participants could not generate a larger set of candidate differentials due to its difficulty.

A graph of different colored bars

Description automatically generated

*Figure 10: Proportion of participants who use each type of reasoning strategy for each condition/case, with the overall proportions across all cases shown by the rightmost bars. The strategies shown are: Hypothetico-Deductive (where multiple differentials are considered simultaneously, orange), Pattern Recognition (where a single differential is considered in turn, blue), Scheme-Induced (where participants evaluate pathophysiological systems as causes of patients rather than specific conditions, green) and None (for cases where a clear differential is not mentioned, grey).*

We note, rather unsurprisingly, that we observe a higher number of average Differential Evaluations when cases are correct (M = 5.85, SD = 0.38) compared to when they are incorrect (M = 4.34, SD = 0.39). Given our methodology for defining accuracy, participants are more likely to mention a correct differential if they mention more differentials. The procedure used in the previous study for collecting data on which differentials participants were considering at each information stage was not present here and hence we are not able to operationalise accuracy in the same manner as before. While we look at which differentials are mentioned, we cannot observe how participants weigh up differentials against each other in the same way as in the first study.

To connect the results of this study to those of Study 1, we break down the same dependent variables (as operationalised in that study) by reasoning strategy. We do not apply statistics to this study due to the lower sample size. We first categorise each of the 6 cases as having a ‘dominant’ reasoning strategy based on which was utilised the most across participants. Through this process, we categorise three conditions as HD (AD, UC, GBS), three conditions as PR (MTB, TTP, TA). The proportions of participants who use each reasoning strategy for each condition can be viewed in Figure 10. We then compare the individual case classifications of strategy to this reasoning strategy that is most commonly used for that medical condition. Table 2 shows how dependent variables are affected by reasoning strategy. We find that the amount of information seeking was fairly consistent across reasoning strategy, but that PR cases were associated with higher value in information seeking. In order to derive informational value, we used the same values of each piece of information for each case that were derived in Study 1. This higher informational value does not translate into higher accuracy for PR cases, though we should note that the manner in which accuracy was defined for this study limits the analysis only to statements made out loud of specific conditions rather than formally recorded differentials as we did in Study 1. In order to formally replicate this finding with the larger dataset, we use the cases from this study and the coding of strategies to apply the same coding to our online dataset from Study 1.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | **Strategy** | **No. of Cases** | **Accuracy** | **Information Amount** | **Information Value** | **Initial Diagnoses** | **Change in Confidence** |
| Think-Aloud (Study 2, training) | HD | 43 | 0.93 | 0.79 | 2.77 | - | 0.19 |
| PR | 29 | 0.66 | 0.78 | 3.38 | - | 0.20 |
| SI | 18 | 0.67 | 0.78 | 2.55 | - | 0.25 |
| Online (Study 1, test) | HD | 183 | 0.46 | 0.63 | 2.15 | 3.37 | 0.16 |
| PR | 128 | 0.39 | 0.50 | 2.35 | 2.84 | 0.14 |
| SI | 199 | 0.39 | 0.66 | 2.42 | 3.28 | 0.19 |

*Table 4: Dependent variables from both studies broken down by the reasoning strategy assigned via a multinomial classifier. We note that Accuracy is defined differently in each study due to participants in Study 2 not providing an explicit written list of differentials.*

Reasoning Strategies in Study 1 Dataset

In order to apply reasoning strategies to the data from Study 1, we train a classifier using penalised multinomial regression to classify cases as HD, PR or SI using the cases from the think aloud study (with Leave One Out Cross Validation). The input parameters for the classifier are the 29 pieces of information as binary predictors (similar to the approach depicted in Figure 7) and the cases’ condition. In other words, the cases from the think-aloud study make up the training data for the classifier whilst the cases from the larger online study is the test dataset. The classifier was implemented using R’s *nnet* package (version 7.3-19). The testing data is then labelled with predicted testing strategies using R’s *predict* function. We note that the training data was initially labelled with reasoning strategies using the think-aloud utterances and thus is separated from the information sought during the case.

We show a breakdown of cases by their coded reasoning strategy in Table 4. We now look to compare our key dependent variables by strategy, in particular comparing PR and HD cases. In line with our expectations based on the definitions of HD and PR reasoning approaches, we find that HD cases are associated with more differentials being considered (M = 3.37, SD = 1.64) average when compared to PR cases (M = 2.84, SD = 1.58) and find evidence of a difference between the two via a Welch Two Sample t-test (t = 2.89, MDiff = 0.53, p = .004). We find that PR cases are associated with higher informational value (M = 2.35, SD = 1.07) when compared to HD cases (M = 2.15, SD = 1.32) (t = 1.48, MDiff = 0.20, p = .14). However we do find evidence of higher amounts of information seeking for HD cases (M = 0.63, SD = 0.21) when compared to PR cases (M = 0.50, SD = 0.21), (t = 5.28, MDiff = 0.13, p < .001). Overall, this indicates that PR reasoning were associated with lower but more selective information seeking when compared to HD reasoning.

We hypothesised that an interaction with reasoning strategy is associated with accuracy on the task. This is because a single reasoning strategy is considered unlikely to be more accurate for all cases. As indicated by Figure 10, different patient conditions seem to result in varying reasoning strategies being utilised, which begs the question of what properties of a condition contribute to changes in strategy and in accuracy. One possibility is that reasoning strategy interacts with the diagnostic uncertainty of a case (i.e. the breadth of conditions that a patient could have given their current symptoms and history, with some conditions presenting in a more apparent way than others), as captured by the number of initial differentials reported by participants. To test this hypothesis, we fit a linear model to predict accuracy with an interaction between the number of initial diagnoses and reasoning strategy. The interaction regression lines are plotted below in Figure 11.

A graph with lines and numbers

Description automatically generated with medium confidence

*Figure 11: Fitted regression line for model to explore the interaction between the number of initial diagnoses/differentials and reasoning strategy (as per our multinomial classifier) for the online dataset from Study 1.*

**Discussion**

In the quantitative results shown in Table 4, we firstly find that in our think-aloud study, PR cases were associated with the highest informational value (despite fairly constant amounts of information seeking across reasoning strategies). SI cases were associated with the highest increase in confidence across the information stages. Whilst the latter finding was similar in the online study, PR cases saw higher informational value than HD cases in this dataset. In accordance with our definitions of reasoning strategies, we find that PR cases were associated the fewest differentials of the three strategies. These findings indicate that different reasoning strategies result in behavioural differences when performing diagnoses, both in terms of information seeking and weighing up differentials.

We also find evidence of an interaction effect between reasoning strategy and the number of initial diagnoses on accuracy. Intuitively, it makes sense to broaden or narrow differentials based on the number of differentials being considered. Given that reasoning strategies differ by a function of the case/condition, it might be that the case-level factor affected reasoning strategy is how ‘apparent’ the underlying condition of the patient is based on the initial patient presentation/history. We operationalise this as the number of initial differentials, which captures how clear the patient’s condition for a given participant based on the patient history. In the case of the interaction depicted in Figure 14, we find that reasoning strategy and the number of initial differentials interact. With a lower number of initial differentials, participants exhibited increased diagnostic accuracy by broadening their consideration of differentials via a hypothetico-deductive process (i.e. “what else could be causing these symptoms?”). With a higher number of initial differentials, higher diagnostic accuracy was found by narrowing differentials via a pattern recognition process (i.e. “which of these differentials does this patient most resemble?”). Whilst past work has tended on focusing on designing cognitive interventions that aim to fit all diagnostic scenarios, this result indicates that a flexibility in reasoning strategy based on the patient case is key for increased accuracy. Future research should hence focus on prompting the right reasoning strategy based on the initial patient presentation.

**General Discussion**

We can consider both studies together to provide a nuanced discussion of the diagnostic process among medical students. We find that information seeking patterns and evaluation of differentials during the diagnostic process contribute to diagnostic accuracy. When students generated a greater number of differentials from a patient history, they sought a greater amount of information. We then observe an association between information seeking and confidence, but not with accuracy. Instead, accuracy was characterised by more selective information seeking during the diagnostic process. This is important to note as it demonstrates that being selective in information seeking is a better marker of performance and giving a lower ability participant all available information does not necessarily translate into accurate diagnoses even though it increases diagnostic confidence (Gruppen, Wolf & Billi, 1991).

This has interesting implications for medical practice, as the ordering of unneeded tests or patient examinations may not contribute to better decisions and is not cost effective. Given the constraints within most hospitals and healthcare to obtain certain tests, being selective with information seeking is already a necessity and results from this study seem to show evidence that it is also a good marker of diagnostic performance. There has been increased research on overtesting, such as requesting costly imaging scans when they may not be medically necessary (Carpenter, Raja & Brown, 2015). ‘Overtreatment’ has been estimated to cost the US healthcare system between 158 and 226 billion dollars in 2011 (Berwick & Hackbarth, 2012). Seeking more information during the task made students more confident but not more accurate, which is important to note as it corresponds with previous findings from the cognitive psychology literature (Ko, Feuerriegel, et al., 2022).

The finding of evidence for an interaction between strategy and the number of diagnoses with regards to accuracy is an interesting one for future medical education. Past work that aim to teach diagnostic reasoning or administer cognitive interventions/aids has tended to assume that a single aid can be optimal for all types of patient cases. However, this results hints at the fact that reasoning strategies’ effects on accuracy depend on the initial diagnostic uncertainty associated with the case. In particular, PR seems to result in lower accuracy for fewer initial diagnoses and higher accuracy for more initial diagnoses. This makes intuitive sense when considering how reasoning strategies relate to reducing or expanding the space of possible diagnoses. For instance, if a clinician has a large set of possible differentials in mind from the initial patient presentation, they should narrow their range of possibilities using a pattern recognition approach (“which of these conditions does this most match?”). Conversely, if a clinician is struggling to bring multiple differentials to mind, they should broaden their thinking to consider more conditions using a hypothetico-deductive approach (“what other conditions should I be concerned about?”). This account of the results is bolstered by our operationalisation of accuracy, whereby participants are more accurate by not only considering the correct condition but also considering it as highly likely amongst the considered differentials.

Coderre et al. (2003) found the pattern recognition was utilised more as clinicians increased in experience. On the one hand, this makes sense given that having more experience of disease presentations would improve a diagnostician’s ability to match symptoms to a condition. However, as alluded to by students in this study, knowledge and experience brings with it the ability to generate more differentials than a less experienced clinician. One cannot adopt a hypothetico-deductive reasoning process, whereby multiple differentials are considered and then eliminated, if the clinician lacks sufficient knowledge to generate a set of differentials based on the observed patient. This may be where the complexity/difficulty of the case has a bearing on reasoning process too, whereby harder cases are harder because one cannot easily generate differentials for them. However, the inverse could also be true, whereby a set of conflicting symptoms may cast a wider net of potential differentials that are more challenging to narrow down. As we noted in the online study, the number of initial differentials has an impact on information seeking behaviour, but as we explain here, differentials are themselves a result of a particular reasoning strategy. Ascertaining the exact interaction between reasoning strategy, case difficulty and differential evaluation is hence important for us to focus on in the following study, as it informs how diagnosis is characterised as a cognitive process and how cognitive interventions are designed to aid the process.

We can also observe that reasoning strategies may in turn bring with them differences in information seeking patterns. The choice of information or tests within the diagnostic has been understudied to date given its role in real-world clinical settings. Our results in Study 1 indicate that information seeking patterns are associated with accuracy, specifically around greater selectivity and less variability in information seeking. This relates well to real-world clinical decisions where information and tests can require sizeable time and even other staff or technology to request, as mentioned earlier. Hence, in a setting where all possible information is not always readily available to clinicians, being selective is advantageous. In addition, being more standardised in information seeking can also make comparisons between patients easier given that the information being compared is more similar. We can assume that a big part of gaining medical experience is by using past patient cases personally experienced by the clinician and then drawing upon that experience for a new incoming patient. This may explain the findings of Coderre et al. (2003) that pattern recognition was utilised more with experience: because experienced clinicians have more past patients to draw patterns from.

Study Strengths

We note a number of strengths of both of these studies. To our knowledge, this is the first research of this kind to use both a mixed-methods approach to understand the cognitive underpinnings of diagnosis as a decisional process. Our paradigm also emulates diagnosis in a manner that is not simply a single decision. When creating a task that emulates diagnosis, we in a sense conceptualise what diagnosis looks like in a fairly static manner, when really diagnosis is a more fluid and nebulous process in medicine. In our study, diagnosis is modelled as a process that develops over time with more information and is constantly shifting doctors’ thought processes. In particular, we note linking diagnostic accuracy, confidence, information seeking and differential generation has not been attempted in prior work and should be considered for future work to consider a more complete study of diagnostic decisions. Future work could build upon on this work to investigate more flexible and open information seeking by emulating naturalistic decision making processes. This can be used to investigate how contextual limitations on information seeking impact confidence and accuracy (such as time pressures or testing being unavailable). The use of a think-aloud paradigm also brings a number of strengths by allowing for an ongoing recording of medical students’ thought processes, again showing the dynamic, evolving nature of the diagnoses. While participants varied in terms of their ability to verbalise their thoughts, it provides a clear access to how they approaching their decisions that would otherwise be difficult to determine in real-time. We therefore encourage future researchers to consider think-aloud methodologies in their work.

Limitations

We note that the use of a think-aloud methodology brings with it a couple of limitations. Firstly, participants may behave differently to how they otherwise would, given that they are being observed and recorded by a researcher. Hence, there may be a tendency toward medical students behaving in a manner that they believe to be judged as better by others, such as being thorough in their information seeking and differential evaluation. Relatedly, we found that medical students naturally differed from one another in terms of the amount of verbalising they did during the task, which could be related to differences in verbalisation skills (van Somersen, Barnard & Sandberg, 1994). By not explicitly asking students for their diagnostic differentials as we did in Study 1 (and minimising the amount of input that the researcher had during the task), we are constrained to analysing only what students say out loud. Given that some students do not verbalise their thoughts as naturally, we may not aware of the aspects of their thought process that they did not verbalise. For example, participants may not explicitly say out loud that a differential is no longer under consideration when in actual fact it has been dropped from their thought process, leading to a lower number of removed differentials as we observed in the data. Similarly, participants may have multiple differentials in mind but some may be subconsciously considered too unlikely to be even worth mentioning. This would then contribute to fewer overt instances of a hypotethico-deductive reasoning process as differentials are underreported. Future work should utilise more structured methods for eliciting clinicians’ thought process during diagnoses in order to ensure accurate reporting of differentials in a more naturalistic, evolving manner. We also could have recruited a larger sample in order to gain a better range of participants and reasoning strategies, increasing the power of our analyses, as well as participants from differing experience levels.

Implications for Medical Practice and Education

There is real value in teaching metacognition and uncertainty within medical education (Royce, Hayes, & Schwartzstein, 2019), such as with the use of cognitive aids (Chew, Durning & Van Merriënboer, 2016, Ely, Graber & Croskerry, 2011), especially given that doctors can be reticent to express their uncertainty (Katz, 1984). A more structured aid is needed, as simply looking at a case for a second time may not be sufficient to improve diagnostic accuracy (Monteiro et al., 2015) and current cognitive forcing strategies have not been found to be effective enough (Sherbino et al., 2014). The reason for this might that past distinction between System 1 (automatic, quick) and System 2 (deliberate, slow) thinking for prompting diagnostic reasoning may be overly simplistic, in that one solution may not fit all possible cases and all clinicians. Future work should focus on understanding when certain reasoning styles and which cognitive aids may be more useful for a given clinical situation. In particular, our findings indicate that cognitive aids should prompt reasoning strategies based on how clear a patient presentation is. This can be thought of as whether the patient’s initial symptoms suggest a narrow or wide range of differentials, which is likely a combination of the clinician’s knowledge and case’s complexity.

Past work on cognitive interventions have not tended to focus on prompting appropriate information seeking, and we show here that different facets of information seeking contribute uniquely to both confidence and accuracy. While the most relevant information that should be afforded to clinicians will differ depending on the medical discipline, interventions can focus on standardising which is the most valuable information to be presented to clinicians in the first place. This could not only improve diagnostic accuracy but ensure more appropriate expressions of confidence and uncertainty by reducing a tendency toward overtesting. We emphasise that such recommendations are highly dependent and variable depending on the specific medical context, but this acts an important facet for medical education to consider around how seeking information relates to reasoning styles and how important these non-technical skills are to integrate into the educational context of medicine.

**References**

1. Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. Cognition, 146, 377-386.
2. Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. The American journal of medicine, 121(5), S2-S23.
3. Berwick, D. M., & Hackbarth, A. D. (2012). Eliminating waste in US health care. Jama, 307(14), 1513-1516.
4. Bhise, V., Meyer, A. N., Singh, H., Wei, L., Russo, E., Al-Mutairi, A., & Murphy, D. R. (2017). Errors in diagnosis of spinal epidural abscesses in the era of electronic health records. The American Journal of Medicine, 130(8), 975-981.
5. Bilalić, M., McLeod, P., & Gobet, F. (2008). Expert and “novice” problem solving strategies in chess: Sixty years of citing de Groot (1946). Thinking & Reasoning, 14(4), 395-408.
6. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.
7. Carpenter, C. R., Raja, A. S., & Brown, M. D. (2015). Overtesting and the downstream consequences of overtreatment: implications of “preventing overdiagnosis” for emergency medicine. Academic Emergency Medicine, 22(12), 1484-1492.
8. Chapman, G. B., Bergus, G. R., & Elstein, A. S. (1996). Order of information affects clinical judgment. Journal of Behavioral Decision Making, 9(3), 201-211.
9. Cheng, J. T., Anderson, C., Tenney, E. R., Brion, S., Moore, D. A., & Logg, J. M. (2021). The social transmission of overconfidence. Journal of Experimental Psychology: General, 150(1), 157.
10. Chew, K. S., Durning, S. J., & Van Merriënboer, J. J. (2016). Teaching metacognition in clinical decision-making using a novel mnemonic checklist: an exploratory study. Singapore medical journal, 57(12), 694.
11. Chi, M. T., Glaser, R., & Farr, M. J. (Eds.). (2014). The nature of expertise. Psychology Press.
12. Clayton, D. A., Eguchi, M. M., Kerr, K. F., Miyoshi, K., Brunyé, T. T., Drew, T., ... & Elmore, J. G. (2023). Are Pathologists Self-Aware of Their Diagnostic Accuracy? Metacognition and the Diagnostic Process in Pathology. Medical Decision Making, 43(2), 164-174.
13. Coderre, S., Mandin, H. H. P. H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. Medical education, 37(8), 695-703.
14. Crowley, R. S., Legowski, E., Medvedeva, O., Reitmeyer, K., Tseytlin, E., Castine, M., ... & Mello-Thoms, C. (2013). Automated detection of heuristics and biases among pathologists in a computer-based system. Advances in Health Sciences Education, 18, 343-363.
15. Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1989). The eyewitness, the expert psychologist, and the jury. Law and Human Behavior, 13(3), 311-332.
16. de Groot, A. D. (1946). Het denken van den schaker. Amsterdam: Noord Hollandsche.
17. Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. Psychological science, 29(5), 761-778.
18. Ely, J. W., Graber, M. L., & Croskerry, P. (2011). Checklists to reduce diagnostic errors. Academic Medicine, 86(3), 307-313.
19. Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychological review, 124(1), 91.
20. Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., ... & Elstein, A. S. (2005). Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. Journal of General Internal Medicine, 20, 334-339.
21. Frotvedt, T. F., Bondevik, Ø., Seeligmann, V. T., & Sætrevik, B. (2020). Primacy, Congruence and Confidence in Diagnostic Decision-Making.
22. Greig, P. R., Higham, H., & Vaux, E. (2015). Lack of standardisation between specialties for human factors content in postgraduate training: an analysis of specialty curricula in the UK. BMJ quality & safety, 24(9), 558-560.
23. Gruppen, L. D., Wolf, F. M., & Billi, J. E. (1991). Information gathering and integration as sources of error in diagnostic decision making. Medical Decision Making, 11(4), 233-239.
24. Hall, K. H. (2002). Reviewing intuitive decision‐making and uncertainty: the implications for medical education. Medical education, 36(3), 216-224.
25. Hautz, W. E., Kämmer, J. E., Hautz, S. C., Sauter, T. C., Zwaan, L., Exadaktylos, A. K., ... & Schauber, S. K. (2019). Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room. Scandinavian journal of trauma, resuscitation and emergency medicine, 27(1), 1-12.
26. Hémon, B., Michinov, E., Guy, D., Mancheron, P., & Scipion, A. (2020). Speaking up about errors in routine clinical practice: a simulation-based intervention with nursing students. Clinical Simulation in Nursing, 45, 32-41.
27. Higgs, J., Jones, M. A., Loftus, S., & Christensen, N. (2008). Clinical reasoning in the health professions E-book. Elsevier Health Sciences.
28. Higham, H., Greig, P. R., Rutherford, J., Vincent, L., Young, D., & Vincent, C. (2019). Observer-based tools for non-technical skills assessment in simulated and real clinical environments in healthcare: a systematic review. BMJ Quality & Safety, 28(8), 672-686.
29. Jaspan, O., Wysocka, A., Sanchez, C., & Schweitzer, A. D. (2022). Improving the relationship between confidence and competence: implications for diagnostic radiology training from the psychology and medical literature. Academic Radiology, 29(3), 428-438.
30. Jonassen, D. H. (1997). Instructional design models for well-structured and III-structured problem-solving learning outcomes. Educational technology research and development, 45(1), 65-94.
31. Kämmer, J. E., Schauber, S. K., Hautz, S. C., Stroben, F., & Hautz, W. E. (2021). Differential diagnosis checklists reduce diagnostic error differentially: a randomised experiment. Medical education, 55(10), 1172-1182.
32. Katz, J. (1984). Why doctors don't disclose uncertainty. Hastings Center Report, 35-44.
33. Ko, Y. H., Feuerriegel, D., Turner, W., Overhoff, H., Niessen, E., Stahl, J., ... & Bode, S. (2022). Divergent effects of absolute evidence magnitude on decision accuracy and confidence in perceptual judgements. Cognition, 225, 105125.
34. Kostopoulou, O., Rosen, A., Round, T., Wright, E., Douiri, A., & Delaney, B. (2015). Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. British Journal of General Practice, 65(630), e49-e54.
35. Kostopoulou, O., Russo, J. E., Keenan, G., Delaney, B. C., & Douiri, A. (2012). Information distortion in physicians’ diagnostic judgments. Medical Decision Making, 32(6), 831-839.
36. Kovacs, R. J., Lagarde, M., & Cairns, J. (2020). Overconfident health workers provide lower quality healthcare. Journal of Economic Psychology, 76, 102213.
37. Kuipers, B., & Kassirer, J. P. (1984). Causal reasoning in medicine: analysis of a protocol. Cognitive Science, 8(4), 363-385.
38. Lawton, R., Robinson, O., Harrison, R., Mason, S., Conner, M., & Wilson, B. (2019). Are more experienced clinicians better able to tolerate uncertainty and manage risks? A vignette study of doctors in three NHS emergency departments in England. BMJ Quality & Safety, 28(5), 382-388.
39. LeBlanc, V. R., Brooks, L. R., & Norman, G. R. (2002). Believing is seeing: the influence of a diagnostic hypothesis on the interpretation of clinical features. Academic Medicine, 77(10), S67-S69.
40. Ly, D. P., Shekelle, P. G., & Song, Z. (2023). Evidence for anchoring bias during physician decision-making. JAMA internal medicine, 183(8), 818-823.
41. McLaughlin, K., Rikers, R. M., & Schmidt, H. G. (2008). Is analytic information processing a feature of expertise in medicine?. Advances in Health Sciences Education, 13, 123-128.
42. McGlynn, E. A., McDonald, K. M., & Cassel, C. K. (2015). Measurement is essential for improving diagnosis and reducing diagnostic error: a report from the Institute of Medicine. Jama, 314(23), 2501-2502.
43. Meyer, A. N., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians’ diagnostic accuracy, confidence, and resource requests: a vignette study. JAMA internal medicine, 173(21), 1952-1958.
44. Monteiro, S. D., Sherbino, J., Patel, A., Mazzetti, I., Norman, G. R., & Howey, E. (2015). Reflecting on diagnostic errors: taking a second look is not enough. Journal of general internal medicine, 30, 1270-1274.
45. Morgan, P. J., & Cleave‐Hogg, D. (2002). Comparison between medical students' experience, confidence and competence. Medical education, 36(6), 534-539.
46. Newell, A., & Simon, H.A. (l972). Human problem solving. Englewood Cliffs. NJ: Prentice-Hall.
47. Orasanu, J., & Connolly, T. (1993). The reinvention of decision making. Decision making in action: Models and methods, 1, 3-20.
48. Payne, J. W. (1994). Thinking aloud: Insights into information processing. Psychological Science, 5(5), 241-248.
49. Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. Nature, 461(7261), 263-266.
50. Redelmeier, D. A., & Shafir, E. (2023). The Fallacy of a Single Diagnosis. Medical Decision Making, 43(2), 183-190.
51. Restrepo, D., Armstrong, K. A., & Metlay, J. P. (2020). Annals clinical decision making: avoiding cognitive errors in clinical decision making. Annals of internal medicine, 172(11), 747-751.
52. Royce, C. S., Hayes, M. M., & Schwartzstein, R. M. (2019). Teaching critical thinking: a case for instruction in cognitive biases to reduce diagnostic errors and improve patient safety. Academic Medicine, 94(2), 187-194.
53. Schiff, G. D., Hasan, O., Kim, S., Abrams, R., Cosby, K., Lambert, B. L., ... & McNutt, R. A. (2009). Diagnostic error in medicine: analysis of 583 physician-reported errors. Archives of internal medicine, 169(20), 1881-1887.
54. Schoenherr, J. R., Waechter, J., & Millington, S. J. (2018). Subjective awareness of ultrasound expertise development: individual experience as a determinant of overconfidence. Advances in Health Sciences Education, 23, 749-765.
55. Sherbino, J., Kulasegaram, K., Howey, E., & Norman, G. (2014). Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: a controlled trial. Canadian Journal of Emergency Medicine, 16(1), 34-40.
56. Van Someren, M., Barnard, Y. F., & Sandberg, J. (1994). The think aloud method: a practical approach to modelling cognitive. London: AcademicPress, 11(6).
57. Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. Acta psychologica, 50(2), 179-197.
58. Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly journal of experimental psychology, 12(3), 129-140.
59. Yang, H., & Thompson, C. (2010). Nurses’ risk assessment judgements: A confidence calibration study. Journal of Advanced Nursing, 66(12), 2751-2760.
60. Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. Journal of Experimental Social Psychology, 33(4), 345-366.